Course Proposal for ESSLLI 2015

a. Personal information

Stepan Kuznetsov, C. Sc. (Ph. D.)
Dept. of Mathematical Logic, Steklov Mathematical Institute, Russian Academy of Sciences
Gubkina str. 8, Moscow 119991, Russia
e-mail: `sk@mi.ras.ru`
homepage: `http://www.mathnet.ru/php/person.phtml?&personid=72238&option_lang=eng`

b. General proposal information

"Lambek categorial grammars"
INTRODUCTORY

c. Contents information

Abstract

In this course we aim to describe Lambek calculus and how it is used to formalise syntax and semantics of fragments of natural languages (mainly English) and give a survey of the principal mathematical results concerning this calculus. The course should be interesting for logic students eager to understand how mathematical logic ideas can work in linguistics and maybe, vice versa, for students from the linguistic realm.

Motivation and Description

Lambek calculus was introduced in 1958 for describing natural language syntax using categorial grammars. This calculus uses syntactic types built from primitive ones using three binary connectives: multiplication (actually rarely needed in practice, but theoretically useful), left and right division. These connectives can very naturally interpreted as operations on formal languages: multiplication stands for concatenation, and B/A contains all words which, being concatenated with any word from A, give a word from B ($A \backslash B$ is defined symmetrically). Lambek calculus is sound and complete with respect to this interpretation (Pentus 1998), which makes Lambek calculus interesting from a purely mathematical point of view and suggests it (very plausibly for mathematicians) to be *the* right calculus for categorial grammars. (There are also other algebraic interpretations of Lambek calculus that also enjoy completeness.)

In a categorial grammar, we associate syntactic types to words of the language (maybe several types to one word) and then check derivability of the corresponding sequent. If the sequent is derivable, the sentence is considered to be syntactically correct. All the syntactic information is kept locally for every word inside the categorial dictionary (the correspondence between words and types); the global part—the calculus itself—is invariant. Thus, if a word has a very non-trivial syntactic behaviour, but doesn't occur in the sentence being parsed, the parsing procedure will remain easy. Also this means that a grammar can be grown by adding new words with their new syntactic features leaving the original fragment untouched.

We have just defined the notion of categorial grammar in a weak sense: the formalism just gives the answer whether a sentence is correct or not. Of course, one wants more from a grammar: if a sentence is actually syntactically correct, the grammar should be able to extract its syntactical structure and the corresponding formal semantics and present it in a suitable way. The common way of dealing with formal semantics in connection with syntax was proposed by Montague. In his approach, the semantic value of a sentence is represented as a typed λ-term. With Lambek categorial grammars, we get Montague-style semantics virtually "out-of-the-box". Being a subsystem of intuitionistic propositional logic, Lambek calculus enjoys the Curry – Howard correspondence. This means that any derivation in the calculus yields a λ-term encoding it; this term can be considered to be the semantic value.

This idea is mainly due to van Benthem. Initially Montague-style semantics was applied to context-free grammars, but, as one can see, its usage for categorial grammars is much more straightforward. However, in the weak sense Lambek grammars and context-free grammars are equivalent (Pentus 1992), i.e., any language generated by a Lambek grammars is context-free and vice versa. This means that Lambek grammars cannot be used (at least without extending the calculus) for describing syntactic phenomena that are known to break context-freeness. Another disadvantage of Lambek's approach is NP-hardness of the derivability problem for Lambek calculus (on the other hand, this becomes a real problem only if the types actually used in the grammar are of high complexity).

Nevertheless, Lambek calculus itself is interesting from the mathematical point of view, and Lambek grammars are a good starting point for exploring ideas of the type-categorial approach to describing natural language syntax.

Tentative Outline

Day 1. Typed λ-calculus and its usage in formal semantics. Formal languages. Basic (Ajdukiewicz-style) categorial grammars: syntax and semantics. The Curry – Howard correspondence: construction of λ-types as logical derivation. Lambek calculus. Lambek categorial grammars.

Day 2. Simple examples: parsing English sentences using categorial grammars. Type raising. More examples: pronouns, coordination, dependent clauses. Limitations of the formalism: non-projective dependencies, unbounded dependencies; examples from languages other than English. Possible workaround strategies.

Day 3. Context-free grammars and Montague-style semantics for them. Connections between Lambek grammars and context-free grammars (Gaifman's, Buszkowski's and Pentus' theorems).

Day 4. Extensions of Lambek calculus. The standard interpretation of Lambek connectives as operations on formal languages. Soundness and (in)completeness results (Buszkowski, Pentus).

Day 5. Algorithmic complexity: NP-completeness of the derivability problem in the Lambek calculus (Pentus); polynomial algorithms for bounded fragments.

Expected level and prerequisites

No particular prior knowledge is needed, but the listeners are supposed to have some very basic background in mathematical logic (e.g., understand such words as "formula", "derivability" etc).

References

[1] B. Carpenter. *Type-logical semantics.* — Cambridge, Mass.: The MIT Press, 1997.

[2] G. Morrill. *Categorial type logic.* — Oxford: Oxford University Press, 2011.

[3] J. Lambek. The mathematics of sentence structure. *American Mathematical Monthly.* **65**, No. 3, 1958. — P. 154–170.

[4] W. Buszkowski. Compatibility of a categorial grammar with an associated category system. *Zeitschrift für mathematische Logik und Grundlagen der Mathematik.* **28**, 1982. — P. 229–238.

[5] W. Buszkowski. The equivalence of unidirectional Lambek categorial grammars and context-free grammars. *Zeitschrift für mathematische Logik und Grundlagen der Mathematik.* **31**, 1985. — P. 369–384.

[6] S. Kuznetsov. Lambek grammars with one division and one primitive type. *Logic Journal of the IGPL.* **20**, No. 1, 2012. — P. 207–221.

[7] M. Pentus. Lambek grammars are context free *Proc. of the 8th Annual IEEE Symposium on Logic in Computer Science.* — IEEE Computer Society Press, Los Alamitos, California, 1993. — P. 429–433.

[8] M. Pentus. Models for the Lambek calculus. *Annals of Pure and Applied Logic,* **75**, No. 1–2, 1995. — P. 179–213.

[9] M. Pentus. Complexity of the Lambek calculus and its fragments. *Advances in Modal Logic,* Vol. **8**, ed. by L. Beklemishev, V. Goranko, and V. Shehtman. — London: College Publications, 2010. — P. 310–329.

The main book covering most ideas in the course (and much more) is [1]. Another good source, particularly for extensions of the formalism, is [2]. Of course, the original Lambek's paper [3] introducing the calculus is worth reading. The theorem that any Lambek language is context-free (Day 3) can be found in [7]. The vice versa direction is originally due to Gaifman and is very well outlined by Buszkowski in [5]. Some clarifications of these results (treatment of the empty word case) are done in [6]. The completeness theorem (Day 4) is only formulated in this course; its detailed proof is in [8]. An easier result for the product-free case is in [4]. Finally, a very good survey of algorithmic results concerning Lambek calculus (Day 5) is given in [9].

d. Practical information

N/A